MULTIMODAL INDEXING OF DIGITAL AUDIO-VISUAL DOCUMENTS: A CASE STUDY FOR CULTURAL HERITAGE DATA

James Carmichael[§], *Martha Larson**, *Jennifer Marlow*[§], *Eamonn Newman*[†], *Paul Clough*[§], *Johan Oomen*[º], *Sorin Sav*[†]

University of Sheffield[§], University of Amsterdam*, Dublin City University[†], Netherlands Institute for Sound and Vision[º]

ABSTRACT

This paper describes a *multimedia multimodal information access sub-system* (MIAS) for digital audio-visual documents, typically presented in streaming media format. The system is designed to provide both professional and general users with entry points into video documents that are relevant to their information needs. In this work, we focus on the information needs of multimedia specialists at a Dutch cultural heritage institution with a large multimedia archive. A quantitative and qualitative assessment is made of the efficiency of search operations using our multimodal system and it is demonstrated that MIAS significantly facilitates information retrieval operations when searching within a video document.

1. INTRODUCTION

The indexing and retrieval of digital audio-visual documents constitutes one of the main objectives of *MultiMatch* (MM)[1], an EU-funded project concerned with providing online access to European cultural heritage (CH) material via a multilingual web-based search engine. The principal motivation here is to develop *information retrieval* (IR) techniques which are specifically designed to meet the research needs of CH professionals. In order to establish user requirements and provide a basis for developing prototypical applications, MultiMatch has identified and secured the cooperation of several CH institutions who serve both as CH content providers and clients/evaluators. In this particular study, we focus on the requirements of a specific MultiMatch client, namely the *Netherlands Institute for Sound and Vision*[2] (hereafter referred to by the shortened version of the organisation's Dutch name, "Beeld en Geluid" or simply "B&G"). One of B&G's public services is the provision – upon request – of copies of audio-visual ~~programmes, particularly~~ television documentaries and ~~newscasts, disseminated in t~~he Dutch mass media.

It is usually the case that B&G clients are not interested in entire video documents but only various segments thereof which are relevant to specific search criteria. Accordingly, this study reports on the prototyping of an online multimedia multimodal document retrieval system that allows users to search within videos for *shot-level* clips that are relevant to their information needs. Individual episodes of television programmes are displayed in the user interface as a series of representative thumbnail images (*key frames*) that act as a visual summary of the video's contents. Any speech data featured on the video document's soundtrack is rendered as a text transcript generated by automatic speech recognition (ASR) technology. Click-and-play functionality allows the user to click on a key frame to initiate video playback from the start of the shot sequence represented by that key frame. Additionally, provision is made for word-level searching of the video's speech transcripts to facilitate the location of relevant shots within the key frame series.

The features offered by this system are designed to meet the specific needs of a group of video professionals and thus optimise their *intra-video* searching and browsing experience. It is to be noted that the combination of techniques described above represent a novel approach to multimedia information retrieval in the cultural heritage domain, an area which is particularly challenging since the presentation formats for CH video documents do not always adhere to any standard conventions as is the case with broadcast news programmes [10] [11].

This paper is organised as follows: the next section reviews previous research and application development in this field; section 3 describes in greater detail the B&G user requirements and the current techniques used by that organization to locate and extract such user-requested audio/video clips; section 4 specifies the design and implementation of the multimodal system under review; sections 5 and 6 present the quantitative and qualitative user evaluation surveys and results. The paper concludes with a discussion of the implications of our prototype system for further developments in the area of audio-visual search systems for multimedia specialists.

[1] http://www.multimatch.org
[2] http://www.beeldengeluid.nl

2. PREVIOUS WORK

In general, previous advances in the design of intra-video browsing systems have been concerned primarily with introducing novel methods for the presentation of a key frame series, examples of such include Yeo & Young's three-tier hierarchical frame sequence display [14], or Boreczky's comic book style layout [2]. These approaches, however, have not attempted to fully exploit the possibilities of using a video document's non-visual data to improve the efficiency of search operations.

Notwithstanding such shortcomings, there have been several attempts to use the automatically extracted speech content from a video's soundtrack for information retrieval purposes. Zhang et al. [15] have developed a multilingual, multimedia information retrieval system that utilises both ASR and machine translation to facilitate search operations on video documents featuring Arabic and/or Chinese language speech content. However, this system – like others incorporating similar approaches – was specifically oriented towards the TRECVid evaluation campaign [4] [11] and therefore focuses on broadcast news video. As Smeaton [11] reminds us, this type of material follows a very specific presentation format and is often visually uninformative (for example, it is often the case that such programmes are dominated by shots of newscasters' faces.) As a result, systems designed for such programme formats may not be optimal for IR analysis of cultural heritage related video (which can be less structured and potentially manifest more background 'noise' on the soundtrack).

Another example of a broadcast news oriented system is the VideoNow news service[3], an online web-based search engine that allows the user to specify alpha-numeric text strings (e.g. "44 and a half billion dollars") as search criteria for browsing through broadcast news programmes. Unfortunately, VideoNow's presentation format is not truly multimodal since it provides no other method of searching: the user must either specify some string of characters or use the application's standard tape recorder style graphical interface to access specific points within the video stream. The application developed during the course of this investigation attempts to redress this functional shortcoming by offering the user two IR modalities – key frame visualisation and speech transcript word search – in order to increase the likelihood of finding a video segment of interest. The following section details the requirements of the targeted user group.

[3] http://videonow.11alive.com/websearch.aspx

3. CASE STUDY: B&G USER REQUIREMENTS

This investigation attempts to identify the IR techniques favoured by CH researchers when procuring excerpts from electronic audio-visual documents for the purposes of creating their own customised presentations (e.g. for the preparation of teaching aids). To this effect, B&G has been selected as representative of a typical multimedia archive. In this case study, therefore, we focus on the search procedures used by CH professionals and members of the general public for locating multimedia material in the B&G digital archive. In determining a typical B&G search scenario, the structured observation method was employed [13]: a series of contextual interviews was conducted at B&G headquarters with four video professionals. Two of these individuals were directly employed by B&G in the customer service section, assisting external clients – such as broadcast agencies – locate requested video clips. The other two individuals were not in B&G's employ but used the archive regularly (one being a freelance video researcher and the other a documentary producer for a Dutch television programme).

The interview questions focused on defining the interviewees' search behaviour, including (i) a general overview of the interviewee's work – as advocated by Marchionini et al [7], Amato et al [1] and Larson et al [6], (ii) the researchers' selection of resources, i.e. reference to web-based search engines and other web sites, (iii) concrete examples of various current and past search processes, the importance of which is discussed in Smeulders et al [12]. The data compiled from these interviews indicated that video documents are located in the B&G archive in a two-stage process. Firstly, all videos which broadly might be of interest are located by submitting text queries via the in-house B&G video browser known as the *Catalogue*. The Catalogue searches the archive using document-level manually annotated metadata, (including series titles, episode titles, production dates and brief thematic descriptions of the video documents' subject material) to return a list of videos relevant to the query. The researcher then chooses any videos from this list that seem particularly relevant. The second stage of the search process then involves identifying appropriate clips *within* a video of interest. Using the Catalogue's graphical interface, the searcher clicks on the appropriate icon representing the video document to view its key frame series. Each key frame is annotated with timestamp information detailing – to the nearest millisecond – the point in the video footage from which the key frame was extracted.

Via visual inspection only, the searcher chooses potentially promising key frames and notes their time stamp information. Subsequently, the appropriate high resolution video cassette is retrieved from a physical storage facility

and the researcher uses a video player to manually cue through the video and preview it at the points indicated by the timestamps of the selected key frames. For those segments of film footage that prove to be relevant when viewed, the client then submits a request for copies of such.

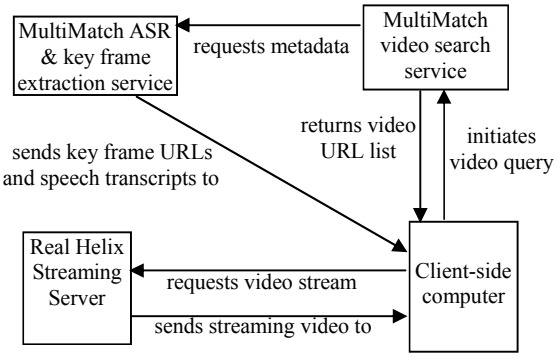
The video-level annotation of the B&G collection is of sufficient scope and detail as to permit the first step in the search process (i.e. the identification of a number of video documents with subject matter meeting the client's search criteria) to be efficiently executed using only the Catalogue's resources. For this reason, the investigation undertaken here focuses solely on the second step of the search process[4]: the location of appropriate shots within a selected video document. Currently, such intra-video searching and browsing often proves an inefficient process since the Catalogue system provides only timestamp information at the shot level; thus when a key frame is located that seems relevant, it is not possible to locate and play the corresponding video segment directly via the Catalogue's search interface, instead it is necessary for the client to make a note of the key frame's shot-start time and then procure the corresponding video cassette in order to access the film segment(s) of interest.

After reviewing the user requirements identified by the interview process and narrowing the domain of investigation to intra-video searching, the following use case scenario has been identified as typical: given specific thematic criteria (e.g. "*Street scenes of 2006 Paris riots*"), the user executes a within-video document search in order to identify clip-length portions of video (i.e. less than 200 seconds). The multimodal retrieval system described in the following section attempts to meet these requirements.

4. MULTIMEDIA RETRIEVAL SYSTEM ARCHITECTURE

The prototype *multimodal information access sub-system* (MIAS) implements shot-level playback and speech content extraction for both streamed and non-streamed video. As mentioned in the previous section, the search operations supported by MIAS are meant to facilitate the second step in the search process, i.e. they start at the point where the user has already identified a particular television programme of interest (using either the MultiMatch search engine or the B&G Catalogue video browser). The data flow schema depicted in Figure 1 represents the operations of the major MultiMatch system components when MIAS responds to a user's query.

Figure 1: MultiMatch Dataflow Schema for Video Retrieval



Once a successful connection has been established with the Real[5] Helix streaming video server and the relevant key frame URL list with corresponding speech transcript text files have been passed to the client computer (enabling a "film strip"-style click-and-play key frame display as depicted in Figure 2), the user is able to visually inspect the key frame sequence while simultaneously perusing the accompanying shot-segmented speech transcripts.

If an apparently relevant key frame and/or speech snippet is encountered, its corresponding footage can be easily retrieved from the parent file since the relevant timestamp information is also provided. The speech and key frame extraction processes which permit such shot-level multimodal indexing are detailed in the section that follows.

4.1 Content-Based Audio-Visual Indexing in MIAS

Shot-level segmentation of video for visual summarisation purposes is a well-established technique [1] [3] [5] [9], with more recent implementations attempting a thematic approach, formatting the key frame display to highlight video segments determined to be particularly relevant [2]. For MIAS, however, a simpler temporally ordered left-to-right display was adopted since this style was more convenient and familiar for the user group participating in this case study. The automatic video document processing protocols are detailed below:

Firstly, shot boundary detection is performed. In ~~this step, the *Cosine Similarity Measure* (CSM) algorithm~~ is used [3] to decompose the videos into their constituent *shots*

[4] The third and final stage in this exercise, namely the process of actually preparing copies of the retrieved document snippets for distribution to the clients, is also beyond the scope of this investigation.

[5] http://www.realnetworks.com/products/media_delivery.html

(a "shot" being defined here as a sequence of frames which are visually similar and therefore probably recorded in a single filming operation). Shot segmentation was effected by comparing the YUV colour histograms (in the uncompressed domain) for every two successive video frames based on the abovementioned CSM metric.

Figure 2: Screenshot of MIAS Interface (with selected video document already loaded)

Given two histogram vectors A and B the cosine measure may be expressed as:

$$CSM(A,B) = 1 - \frac{A \bullet B}{\| A \| \cdot \| B \|} = 1 - \frac{\sum_i a_i \cdot b_i}{\sqrt{\sum_i (a_i)^2 \cdot (b_i)^2}} \quad (1)$$

where a_i , b_i, are the colour components in the histogram vectors and $A \bullet B$ is the dot product. The CSM incorporates the cosine of the angle between the two histogram vectors in a measure that expresses dissimilarity, thus the larger the CSM value, the greater the dissimilarity. A *shot boundary* is detected when the dissimilarity exceeds a predefined threshold. However, the use of only one threshold is not adequate for gradual transitions (e.g. fades and wipes) that typically extend over multiple frames. The values extracted from such gradual shifts can usually be described by a bell-shaped curve where inter-frame dissimilarity increases steadily before peaking and then declining as the transition completes. A double-threshold method is employed [3] to detect these gradual shift patterns.

After shot boundary detection has been completed, the key frame extraction step is performed. During the extraction step, the most representative frame is selected for each shot. The most representative frame is that frame with histogram vector values that best typify the entire set of frames present in that particular shot sequence.

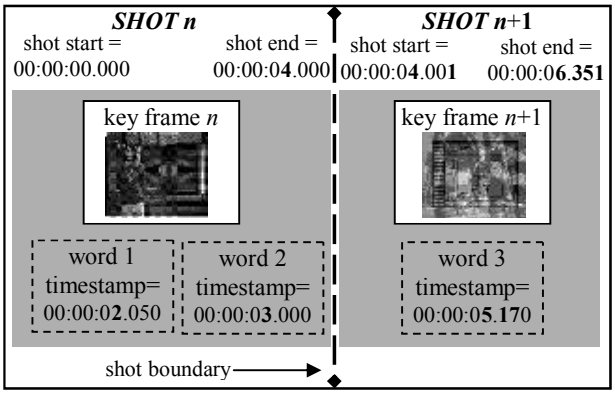
Finally, ASR processing is applied to the audio track, which has been extracted from the video using the "ffmeg" application, an open source tool for video processing.[6] The ASR transcripts were generated by the Dutch language version of Nuance Dragon Naturally Speaking 9 SDK Server Edition used "out of the box"[7]. The emphasis in the prototype was placed on shot-level access and not on optimization of absolute speech transcript quality since it was determined that the recognition rate achieved by the ASR application without any form of speaker or language model adaptation was sufficient for the purposes of evaluating the prototype system. The ASR application outputs a transcription of speech content of the entire television programme, linking each word in the transcript with a time code to indicate when it is spoken. Such word-level time-stamped transcripts are then synchronised (see Figure 3) with the shot boundary time codes output by the key frame extractor. For ease of data transfer between server and client computers, this merged timestamp information is encoded in a single XML-formatted document where all time codes are rendered as offsets from the beginning of the video stream (defined in this instance as the first frame of the video file).

The audio and visual information is synchronised via reference to the time codes of the shot boundaries. Expressing the key frame and speech transcript time codes as offsets serves two functions: (i) it renders possible the implementation of the click-and-play functionality deemed important for intra-video shot selection and browsing; (ii) it allows any user-specified words or short phrases to be used as query items to search the individual soundtracks of the shot-level segments. This implementation of speech/key

[6] http://ffmpeg.mplayerhq.hu/
[7] http://www.nuance.com/audiomining/sdk/

frame synchronisation makes it possible for MIAS to use the shot as the basic unit of retrieval rather than its parent video document. The presentation of this key frame and speech transcript data to the user now merits closer consideration.

Figure 3: Schema of MIAS' Speech Transcript/Key Frame Synchronisation



4.2 User Interface

The MIAS user interface, pictured in Figure 2, includes six components: (i) a text input box in which the user enters query terms; (ii) the main video playback viewing screen; (iii) a standard tape recorder-style console featuring button and slider controls for starting, stopping, pausing etc. video playback; (iv) the left-to-right click-and-play key frame display; (v) a text box in which the segment-specific speech transcripts appear; (vi) a third text box – the lowermost text box of the user interface as shown in Figure 2 – to display all the segments in which the user-defined search terms have been found.

The presentation of the ASR-derived speech transcripts in their entirety (despite any speech recognition errors therein) is meant to allow the user to gain an overall impression of a selected segment's semantic content. Transcript quality, especially in cases with background music or accented speech, may not be sufficient to allow for easy readability. In this context, the transcript acts more as an automatically generated term cloud giving the gist of what was said rather than a verbatim record of the soundtrack's spoken content.

During the development of the prototype, experiments were performed with automatically generated term clouds rather than transcripts. After interacting with both presentation types of speech presentation options, however, the CH professionals and lay users preferred the "raw" speech transcripts since they might contain the occasional phrase which – although not initially specified as a search term – could still prove to be of some interest.

The following section details the evaluation methodology and protocol used to determine the user group's overall impressions of MIAS' usability and fitness for purpose.

5. DESIGN OF USER EVALUATION SURVEYS

The principal objective of this evaluation was to determine the extent to which the MIAS system succeeded in optimising the intra-video search process (as described in section 2). In this context, "search optimisation" may be evaluated using the following criteria:

i. *Efficiency of information access and retrieval*: as demonstrated by several studies [2] [8] [15], the speed of an IR system when returning search results is a significant factor in determining user satisfaction. In the case of the MIAS system, all possible search results for a given intra-video browsing operation are transferred to the client's computer when a specific video document is loaded for playback (see Figure 2); a more appropriate measure of MIAS' efficiency, therefore, would be the time taken for the system to actually commence playback in response to the user clicking on any of the key frames. This playback time lag is used as the principal metric to assess system efficiency.

ii. *Relevance of returned items*: Arguably the most important feature of an IR application is its capacity to distinguish items of relevance in relation to the user's query. The performance of MIAS in this regard will be assessed via the execution of pre-set search tasks.

iii. *Qualitative Assessment of Application Usability*: Issues concerning user satisfaction with the MIAS user interface graphical design will be assessed by way of a questionnaire eliciting qualitative responses from the user.

5.1 Evaluation Tasks and Protocols

A sample of ten individuals (none of whom were members of the original user requirements capture group mentioned in section 3) participated in the system evaluation. Three of these persons were CH professionals, either from B&G or

with a similar background. Due to time and resource constraints, the other individuals were not CH specialists but had some experience interacting with online video (e.g. YouTube, VideoNow) and boasted at least the equivalent level of technical understanding as the B&G employees. In order to quantitatively assess the performance of MIAS, the following evaluation protocol was devised: a thirty-minute video document was selected of which the subject matter was the career and artistic output of the well-known Spanish painter, Pablo Picasso. The evaluators were provided with pre-defined search criteria, expressed as the following phrases: (i) "*Picasso and cubist period*" and (ii) "*Groningen referendum*". The ten participants were then instructed to identify clip-length video segments (i.e. less than 200 seconds) considered relevant to the abovementioned criteria. Furthermore, the evaluators were permitted some latitude in how they interpreted these pre-defined queries, i.e. they could include in their search any word or phrase considered relevant to the theme suggested by the pre-defined search criteria, e.g. when searching for clips pertaining to the "Groningen referendum" theme, an evaluator was at liberty to include the term "Groningen survey" in order to increase the likelihood of locating a segment of interest. This degree of latitude was permitted in order to simulate the process of query refinement, a typical behaviour for query operations of this nature [6] [9].

For the purposes of determining the completeness of the MIAS-supported information retrieval process vis-à-vis finding relevant clips, all of the participants independently viewed the video document in its entirety after finishing the evaluation procedure described above. During this final viewing, each assessor was requested once again to select all video segments which were pertinent to the search criteria. The number and quality of segments in this final selection is then regarded as the "ground truth" relevance measure, i.e. the assessor's ideal choice of snippets against which to compare the initial selection made by the same individual when restricted to searching via MIAS' summarisation techniques. Any change in an evaluator's initial and final selection of clips is expressed as a *shot difference* (SD) measure, whereby the actual shots which comprise the clips selected in the first and second rounds are compared to determine the extent of overlap.

The following example illustrates the computation of the SD measure: after the second viewing of a selected video, an evaluator chooses two clips, each consisting of ten specific shots. In the first viewing (with the assistance of MIAS), the same evaluator had originally selected one clip made up of fifteen shots. If the clip selected during the first round has only five shots which differ from those clips selected in the second round, then the shot difference percentage would be 25.0% (representing the five out of twenty-five shots which differ between the user's selection of clips for the first and second rounds).

In terms of assessing speech transcription accuracy, a count was made of the number of times that MIAS' ASR component correctly identified the five key terms ("Picasso", "cubist", "period", "Groningen" and "referendum") in the video document's soundtrack. Similarly, in an attempt to assess the quality of MIAS' key frame extraction process, every instance of *key frame duplication* – the appearance of two frames depicting virtually identical visual content and originating from the same shot sequence – was also recorded.

In addition to the quantitative assessment procedures described above, the users' qualitative impressions of MIAS' performance was also elicited via a questionnaire which contained the following items (the participants giving their responses in the form of a score from "0" to "5" – ranging from total disagreement to very strong agreement):

(a) *Did you find the speech transcripts helpful when searching for clips (for example, were there occasions when you relied more on the speech transcript than the corresponding key frame when searching)?*

(b) *After reviewing the video in its entirety, is it still your impression that the key frames are adequately representative of the video document's entire film footage?*

(c) *Was the MIAS system's response time and quality of video playback acceptable for your purposes?*

The users' qualitative impressions of MIAS' usability and capacity to optimise the search process – along with a quantitative assessment of the system's performance – are presented in the section that follows.

6. USER EVALUATION RESULTS

In terms of correctness of automatic classification, MIAS' key frame extraction and ASR accuracy would appear to be adequate. For the selected video from which 156 key frames were extracted, only three per cent (5 frames) were duplicates; this percentage compares quite well with state of the art applications in this domain [3].The key terms spoken in the video appeared in the transcript two out of three times. Neither duplicate shots nor dropped key terms had a significant negative effect on the usefulness of MIAS' representation of shot level documents in the retrieval process. Seven out of the ten evaluators (including all three of the CH professionals) did not alter their MIAS-assisted

selection of segments even after viewing the video file in its entirety. For the three evaluators who did make alterations, their differences in shot selection were 3.8%, 12.6% and 19.0% respectively.

Table 1: User Scores (maximum of 5) for the MIAS Evaluation Questionnaire (see section 5.1)

User ID	Q (a)	Q (b)	Q (c)
1	4	4	3
2	4	5	2
3	4	5	3
4	3	3	3
5	4	4	3
6	4	4	2
7	4	4	2
8	2	3	1
9	4	4	1
10	3	3	3
Avg. Score	**3.6**	**3.9**	**2.3**

The users' responses to the qualitative assessment questionnaire (see Table 1) indicate that the provision of speech transcripts played an integral role in locating and identifying relevant clip-length segments. The ten scores of the individual evaluators to the first question of the questionnaire – see section 5.1 (a) – averaged 3.6, with the median score being "4". The comments of one of the evaluators regarding the usefulness of the shot-aligned speech transcripts typify the sentiments of his colleagues: "*...for instance when you see a talking head, the transcript will disclose what this person is talking about*". The key frame's representation of the video document's entire contents was also judged to be satisfactory, with the evaluators according an average rating of 3.9 in response to question 5.1 (b).

The only area where there was a notable level of dissatisfaction concerned the speed of the video playback, as evidenced by the evaluators' mean average rating of 2.3 for question 5.1(c). On some occasions, commencement of video playback – when initiated by clicking on a key frame – was excessively delayed due to poor connectivity with the remote streaming server. These bandwidth problems were, in all likelihood, also responsible for occasional deterioration in playback quality (dropping of frames, pixelation etc.) which reduced user satisfaction with the MIAS application.

Overall, the qualitative and quantitative assessments of the MIAS system conducted in this investigation indicate that the task of selecting relevant clips from some video document is indeed facilitated by presenting the user with both shot-aligned speech transcripts and a series of key frames which comprehensively summarise the document's visual content. There is, however, much scope for further research in the area of improving the ASR accuracy, especially in the case of videos featuring multilingual speech content. Possible techniques for realising such improvements are discussed in the ensuing section.

7. CONCLUSIONS AND FUTURE WORK

As emerged during the evaluation, both key frame duplication and dropped keywords were noticeable in the interface. Although these errors did not have a significant impact on the successful completing of the evaluation task, it is clear that improvement in these areas would make possible even more accurate, and consequently more helpful, video representations. Future work will involve experimentation with concatenation of contiguous shot sequences to eliminate key frame duplicates. An immediate corrective measure – now that the usefulness of speech transcript for shot-level intra-video search has been established – would be to take advantage of the full capacity of the ASR application to adapt to speakers and specialist vocabularies / jargon for CH topic domains. It is clear, however, that such upgrades to the existing system functionality will not in themselves be sufficient to always correctly extract the rich multilingual and multimodal information contained in the 50 CH videos selected for this case study. As mentioned in section 3.1, it is not uncommon for a B&G video's soundtrack to feature multilingual speech; additionally, there is occasionally non-speech acoustic data (e.g. classical music) which may well be of cultural interest. It would therefore be useful to implement some form of automatic language identification application capable of determining if – within a particular soundtrack – a given instance of speech is recognisable as one of the MM-supported languages or otherwise. The idea here is that the language identifier would perform a preliminary "first pass" over the speech data, annotating every encountered speech fragment as being Spanish, Dutch, English, etc. Upon completion of this first pass, it would then be straightforward to select the appropriate language-specific speech recogniser to decode the pre-tagged series of multilingual fragments. Such language recognition capability would thus appreciably increase overall ASR accuracy and, furthermore, could also operate in conjunction with an automatic musical instrument identifier in order to isolate those segments of a soundtrack featuring musical performances of CH interest.

Apart from the abovementioned automatic techniques for ASR techniques, it is envisaged that a user-defined tagging protocol will be defined which would allow

not only the manual correcting of ASR soundtrack speech transcripts but also the introduction of a manual annotation protocol permitting a Flickr-style tagging of speech transcripts. These protocols would enable the indexing of CH-relevant non-speech acoustic data (such as music) for later reference. The full potential of such enriched metadata tagging would only be evident over an extended period of time, i.e. after users would have had sufficient time to insert the metadata given that such activity would be casual and episodic.

Finally, support for content based visual queries for video documents (whereby the search criterion is not a text string but some image in the form of a graphic file) is actively being researched and developed within the MultiMatch project. The MultiMatch search service is already capable of image similarity searches based on low level features (e.g. via the comparison of targeted images' colour histograms) and it is expected that such functionality could be extended to include similarity searches of video key frames.

The MIAS system described in this paper has been designed from a user-centred perspective in order to meet the observed needs of professionals who deal with video material. Initial results indicate that the system helps users to navigate and interact with said content in ways that are currently not available to them. It is envisaged that the novel search techniques made possible by all the abovementioned functionality – both present and proposed – will, in their entirety, serve to realise a new IR paradigm in the cultural heritage domain.

8. ACKNOWLEDGEMENTS

This study has been realised and funded by the MultiMatch EU STREP project (Contract #033104).

9. REFERENCES

[1] Amato, G., Gennaro C., and Savino, P. "Indexing and Retrieving documentary films: managing metadata in the ECHO System", *Proceedings of 4th International Workshop on Multimedia Information Retrieval* (MIR 2002), Paris, France, December 2002.

[2] Boreczky, J., Girgensohn, A., Golovchinksy, G, Uchihashi, S., "An Interactive Comic Book Presentation for Exploring Video", *CHI Letters*, Vol. 2, No. 1, pp. 185-192, 2000.

[3] Calic, J., Sav, S,, Izquierdo, E., Marlow, S., Murphy, N., O'Connor, N., "Temporal Video Segmentation for Real-Time Key Frame Extraction", *International Conference on Acoustics, Speech and Signal Processing ICASSP 2002*, Orlando, Florida, May 2002.

[4] Christel, M., Yan, R., "Merging Storyboard Strategies and Automatic Retrieval for Improving Interactive Video Search", *International Conference on Image and Video Retrieval, Amsterdam*, pp. 486-493, Amsterdam, July, 2007.

[5] Kuo, T.C.T., Lin, Y.B., "Efficient Shot Change Detection on Compressed Video Data", *IEEE Trans. On Circuit and Systems for Video Technology*, pp. 101-108, May, 1996.

[6] Larson, M. , Eickeler, S., Köhler, J., "Supporting Radio Archive Workflows with Vocabulary Independent Spoken Keyword Search", *Proceedings of SIGIR 2007 Workshop Searching Spontaneous Conversational Speech*, pp. 21-28 Amsterdam, July 2007.

[7] Marchionini, G., Geisler, G., "The Open Video Digital Library", *D-Lib Magazine*. Vol. 8, No. 12, December 2002.

[8] Morris, J., Neuwirth, C., Regli, S., Chandok, R., Wenger, G., "Interface Issues in Computer Support for Asynchronous Communication", *ACM Computing Surveys* (CSUR), Vol. 31, No. 11, pp. 1-6, June 1999.

[9] Petrelli, D., Auld, D., Gurrin, C., Smeaton, A., "Retrieving Amateur Video from a Small Collection: Investigating Technical Challenges and User Experience" 9th European Conference on Digital Libraries ECDL 2005, pp. 1-16, Vienna, September 2005.

[10] Renals, S., Abberley, D., Kirby, D., Robinson, T., "Indexing and Retrieval of Broadcast News", *Speech Communication*, Vol. 32, No. 1, pp. 5-20, September 2000.

[11] Smeaton, A., Over, P., Kraaj, W, "Evaluation campaigns and TRECVid", *Proceedings of the 8th ACM International Workshop on Multimedia IR* (MIR 2006), pp. 321-330, California, USA, 2006.

[12] Smeulders, A.W.M., de Jong, F., Worring, M., "Multimedia information technology and video annotation" In:, *Beeld en Geluid*, (Mieke Lauwers, ed.), pp. 4-19 Jaarboek SAP, 2006.

[13] Van Manen, M., *Researching Lived Experience: Human Science for an Action Sensitive Pedagogy*, London, Ontario, Althouse Press, 1990.

[14] Yeo, B., Yeung, M., "Retrieving and Visualizing Video", *Communications of the ACM*, Vol. 40, No. 12, pp. 43-52, December 1997.

[15] Zhang, P., Plettenberg, L, Klavans, J.L., Oard, D.W., Soergel, D., "Task-based interaction with an integrated multilingual, multimedia information system: A formative evaluation" *Proceedings of 2007 Conference on Digital Libraries*, pp. 117-126, Vancouver, June 2007.